Steady-State Distribution Convergence for $GI/GI/1+GI$ Queues in Heavy Traffic

Chihoon Lee[*]
Department of Statistics
Colorado State University
Fort Collins, CO 80523

Amy Ward[†]
The Marshall School of Business
University of Southern California
Los Angeles, CA 90089

April 6, 2015

Abstract

We establish the validity of the heavy traffic steady-state approximation for a single server queue, operating under the FIFO service discipline, in which each customer abandons the system if his waiting time exceeds his generally-distributed patience time. This follows from early results of Kingman when the loading factor approaches one from below, but has not been shown in more generality. We prove the convergence of the steady-state distributions of the offered waiting time process and their moments both under the assumption that the hazard rate of the abandonment distribution is scaled and that it is not scaled. As a consequence, we establish the limit behavior of the steady-state abandonment probability and mean queue-length.

Keywords: Customer Abandonment; Heavy Traffic; Steady-state Convergence

1 Introduction

There have been several papers that have studied queueing systems with customer abandonments. The knowledge of the long time asymptotic behavior (stability) of such systems is of great importance in practice; however, excepting special cases, the models of interest are too complex to analyze directly. The heavy traffic limits often provide tractable, parsimonious, and yet very meaningful approximations. In this paper, we consider a single server queue, operating under the FIFO service discipline, with generally-distributed

[*] E-mail: chihoon@stat.colostate.edu

[†] E-mail: amyward@marshall.usc.edu

patience time (the $GI/GI/1+GI$ queue). To the best of our knowledge, the only results available in the literature in the $GI/GI/1+GI$ setting are process-level convergence results. Our aim is to fill this gap by establishing the convergence of the steady-state distributions/moments of the offered waiting time process for the $GI/GI/1+GI$ queue in heavy traffic. As a consequence, we derive the limit behavior of the steady-state abandonment probability and also the mean steady-state queue-length. These results are necessary to rigorously solve steady-state performance optimization problems in the heavy traffic limit.

Our asymptotic analysis relies heavily on past work that has developed heavy traffic approximations for the $GI/GI/1+GI$ queue, using the offered waiting time process. The offered waiting time process, introduced in Bacelli et al. (1984), tracks the amount of time an infinitely patient customer must wait for service. Its heavy traffic limit when the abandonment distribution is left unscaled is a reflected Ornstein-Uhlenbeck process (see Ward and Glynn (2005)), and its heavy traffic limit when the abandonment distribution is scaled through its hazard rate is a reflected nonlinear diffusion (see Reed and Ward (2008)). However, those results are not enough to conclude that the steady-state distribution of the offered waiting time process converges, which is the key to establishing the limit behavior of the steady-state abandonment probability and mean queue-length. Those limits were conjectured in Reed and Ward (2008), and shown through simulation to provide good approximations. However, the proof of those limits was left as an open question.

When the system loading factor is less than one, since the $GI/GI/1+GI$ queue is dominated by the $GI/GI/1$ queue, the much earlier results of Kingman (1961, 1962) for the $GI/GI/1$ queue can be used to establish the heavy traffic steady-state approximation for the $GI/GI/1+GI$ queue. The difficulty arises because, in contrast to the $GI/GI/1$ queue, the $GI/GI/1+GI$ queue has a steady-state distribution when the system loading factor equals or exceeds 1 (see Bacelli et al. (1984)). Furthermore, the results in the aforementioned papers only provide the convergence of the steady-state distribution, but not its moments.

An informed reader would recall Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009), which establish the validity of the heavy traffic steady-state approximation for a generalized Jackson network, without customer abandonment. The proof of the former paper Gamarnik and Zeevi (2006) relied on certain exponential integrability assumptions on the primitives of the network and as a result a form of exponential ergodicity was established. The latter paper Budhiraja and Lee (2009) provided an alternative proof assuming the weaker square integrability conditions that are commonly used in heavy traffic analysis. Our analysis is inspired by the methodology developed in the latter work Budhiraja and Lee (2009) and we expand their approach to accommodate various customer abandonment settings. The key differences are that (a) the presence of customer abandonments means we must also use the properties of a (nonlinear) generalized regulator map in our analysis; (b) we have to worry about the difference between the offered waiting time process, which does not include customers that will eventually abandon, and the queue-length process, that does; (c) the scaling of the hazard rate distribution leads to a connection between the tail behavior of the limiting hazard function and the number of moments that must be assumed on the inter-arrival and service times to ensure the convergence of the steady-state moments (which is not required to obtain distributional convergence only).

In comparison to results for many-server queues, the process-level convergence result for the $GI/GI/N + GI$ queue in the quality-and-efficiency-driven regime was established in Mandelbaum and Momcilovic (2012) when the hazard rate is not scaled, and in Reed and Tezcan (2012), under the assumption of exponential service times, when the hazard rate is scaled. Neither paper establishes the convergence of the steady-state distributions. That convergence is shown under the assumption that the abandonment distribution is exponential and the service time distribution is phase type in Dai et al. (2014). The question remains open for the fully general $GI/GI/N + GI$ setting.

The remainder of this paper is organized as follows. In Section 2, we set up the model assumptions and recall the known process-level convergence results for the $GI/GI/1+GI$ queue both when the hazard rate distribution is not scaled and when it is. In Section 3, we state our main result, that gives the convergence of the steady-state distribution of the offered waiting time process, and its moments. As a corollary to this result, we obtain the convergence of the steady-state abandonment probability and mean queue-length. Section 4 shows how to obtain bounds on the moments of the scaled state process that are uniform in the heavy traffic scaling parameter (n). We use those bounds in Section 5 to prove our main result.

2 The Model and Known Results

The $GI/GI/1+GI$ model having FIFO service is built from three independent i.i.d. sequences of non-negative random variables $\{u_i, i \geq 1\}$, $\{v_i, i \geq 1\}$, and $\{d_i, i \geq 1\}$, that are defined on a common probability space $(\Omega, \mathcal{F}, I\!\!P)$, and for which $I\!\!E[u_1] = I\!\!E[v_1] = 1$ and $I\!\!E[u_1^2] < \infty, I\!\!E[v_1^2] < \infty$.

We consider a sequence of systems indexed by $n \geq 1$ in which the arrival rates become large and service times small. (By a convention, we will superscript any process or quantity associated with the n-th system by n.) More specifically, assume a sequence of arrival rates indexed by $n \geq 1$ defined as

($\mathbb{A}1$) $\lambda^n \equiv n\lambda$ for some $\lambda > 0$.

The i-th arrival to the system having arrival rate λ^n occurs at time

$$t_i^n = \sum_{j=1}^{i} \frac{u_j}{\lambda^n},$$

and has service time

$$v_i^n = \frac{v_i}{\mu^n},$$

where μ^n is a service rate in the n-th system. A customer abandons without receiving service if processing does not begin by time $t_i^n + d_i$. We let F be the cumulative distribution function of d_1, and assume it is proper (that is, $F(x) \to 1$ as $x \to \infty$).

We impose the standard heavy traffic assumption.

($\mathbb{A}$2) $\sqrt{n}\left(\frac{\lambda^n}{n} - \frac{\mu^n}{n}\right) \to \theta \in (-\infty, \infty)$ as $n \to \infty$.

For example, when

$$\mu^n = n\lambda - \sqrt{n}\theta + o(\sqrt{n}) \text{ for some } \theta \in (-\infty, \infty),$$

so that the mean service times become short as the demand becomes large, ($\mathbb{A}$2) holds. The arrival and service rates are approximately equal for large n because ($\mathbb{A}$1) and ($\mathbb{A}$2) imply

$$\frac{\mu^n}{n} \to \lambda \text{ as } n \to \infty \text{ and } \frac{\lambda^n}{n} = \lambda \text{ for every } n.$$

Finally, the following assumption will be required to ensure each system in the sequence is stable.

($\mathbb{A}$3) The inter-arrival and service time distributions are such that

$$\sup\{x : I\!\!P(v_1 \le \mu^n x) = 0\} - \inf\{x : I\!\!P(u_1 \le \lambda^n x) = 1\} < 0 \text{ for each } n \ge 1.$$

The Offered Waiting Time and Queue-Length Processes

The offered waiting time process, first given in Bacelli et al. (1984), tracks the amount of time an incoming customer at time t has to wait for service. That time depends only upon the service times of the non-abandoning customers already waiting in the queue, that is, those waiting customers whose abandonment time upon arrival exceeds their waiting time. For $t > 0$, the *offered waiting time* process having initial state $V^n(0) = x_0^n$ has the evolution equation

$$V^n(t) = x_0^n + \sum_{j=1}^{A^n(t)} v_j^n \mathbf{1}_{[V^n(t_j^n -) < d_j]} - \int_0^t \mathbf{1}_{[V^n(s) > 0]} ds \ge 0, \tag{1}$$

where

$$A^n(t) \equiv \max\{i : t_i^n \le t\}. \tag{2}$$

The quantity $V^n(t)$ can also be interpreted as the time needed to empty the system from time t onwards if there are no arrivals after time t, and hence it is also known as the workload at time t.

The queue-length process $Q^n(t)$ represents the number of customers that are in the system at time $t > 0$, either waiting or with the server. In contrast to V^n, Q^n includes customers that will eventually abandon but have not yet done so. Fortunately, the number of such customers is small enough that they can be safely ignored in our asymptotic regime, as will be seen. The implication for our analysis is that we can focus on V^n, and obtain results on Q^n from results on V^n.

The steady-state distributions of V^n and Q^n exist uniquely from Bacelli et al. (1984) and/or Lillo and Martín (2001). We let $V^n(\infty)$ and $Q^n(\infty)$ be the random variables having the respective steady-state distributions.

Conventional knowledge states that the queue-size is of order of the square-root of the arrival rate, and so a sample path version of Little's law known as the snapshot principle suggests that

$$V^n(t) \approx \frac{Q^n(t)}{\lambda^n} \approx \frac{1}{\sqrt{n}} \frac{1}{\lambda}, \quad \text{for } t > 0.$$

The implication of the offered waiting time process becoming small is that it is only the most impatient customers that abandon. More specifically, assume $F'(0)$ is well-defined and the initial condition satisfies $\sqrt{n}x_0^n \to x$ as $n \to \infty$. Let (V, L) be the unique solution to the reflected stochastic differential equation

$$\begin{array}{l} V(t) = x + \sigma W(t) + \frac{\theta}{\lambda}t - F'(0) \int_0^t V(s)ds + L(t) \geq 0 \\ \text{subject to: } L \text{ is non-decreasing, has } L(0) = 0 \text{ and } \int_0^\infty V(s)dL(s) = 0, \end{array} \tag{3}$$

where $\{W(t) : t \geq 0\}$ denotes the one-dimensional standard Brownian motion, and the infinitesimal variance parameter is

$$\sigma^2 \equiv \lambda \operatorname{var}(u_1) + \lambda^{-1} \operatorname{var}(v_1).$$

(See the paragraph below (19) for a derivation.) Then, Theorems 1 and 3 in Ward and Glynn (2003) show that

$$\sqrt{n}V^n \Rightarrow V \text{ and } \frac{Q^n}{\sqrt{n}} \Rightarrow \lambda V, \text{ in } D(I\!R_+, I\!R) \text{ as } n \to \infty, \tag{4}$$

where $D(I\!R_+, I\!R)$ is the space of right-continuous functions with left limits, endowed with the Skorokhod J_1-topology (see, for example, Billingsley (1999)), and the symbol "$\Rightarrow$" stands for the weak convergence.

The convergence in (4) motivates approximating the scaled steady-state distributions for V^n and Q^n, and their steady-state moments, using the steady-state distribution of V, and its moments. This is convenient because there is the analytic expression for $\mathbb{E}[V(\infty)]$ in Browne and Whitt (1995), which gives

$$\mathbb{E}[V(\infty)] = x + \frac{\theta}{\lambda F'(0)} + \frac{\sigma}{\sqrt{2F'(0)}} h_Z \left(\frac{-\theta}{\lambda \sigma} \sqrt{\frac{2}{F'(0)}} \right),$$

where $h_Z(\cdot)$ is the hazard rate function of the standard normal distribution. When the abandonment distribution is exponential with rate γ, $F'(0) = \gamma$, and the parameter γ fully characterizes the abandonment distribution. Otherwise, there is the potentially undesirable feature that the approximation depends only on the value of the abandonment density at 0.

Incorporating the Entire Abandonment Distribution

We apply the hazard rate scaling in Reed and Ward (2008), which captures the full abandonment distribution. To do this, we must allow the abandonment distribution to change with n. Specifically, we replace the sequence $\{d_i, i \geq 1\}$ with $\{d_i^n, i \geq 1\}$, and let F^n be the cumulative distribution of d_1^n, and h^n its associated hazard rate function. Note that when F^n has support on $[0, \infty)$, then the relationship

$$F^n(x) = 1 - \exp\left(- \int_0^x h^n(u) du \right), \quad x \geq 0, \tag{5}$$

holds, and we assume that h^n is a nonnegative and continuous function on $[0, \infty)$.

To have an approximation in which the entire abandonment distribution appears, it is necessary to modify the diffusion in (3). For that, we begin by considering a customer arrival at time $s > 0$, that finds the offered waiting time to be $V^n(s)$. The probability that would-be customer abandons is

$$F^n(V^n(s)) = 1 - \exp\left(- \int_0^{V^n(s)} h^n(u) du \right),$$

from (5). The recollection that waiting times are of order $1/\sqrt{n}$ suggests the change of variable $u = \sqrt{n}v$, because then

$$F^n(V^n(s)) = 1 - \exp\left(\frac{-1}{\sqrt{n}} \int_0^{\sqrt{n}V^n(s)} h^n\left(\frac{v}{\sqrt{n}} \right) dv \right),$$

and we expect the upper limit of integration to be convergent. Next, a Taylor series expansion suggests the approximation

$$F^n(V^n(s)) \approx \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}V^n(s)} h^n\left(\frac{v}{\sqrt{n}} \right) dv.$$

The assumption that

$$h^n(x) = h(\sqrt{n}x) \tag{6}$$

for some given hazard rate function h implies that the scaled instantaneous loss rate for these would-be customers is

$$\sqrt{n}F^n\left(V^n(s)\right) \approx \int_0^{V(s)} h(v)dv, \tag{7}$$

where $V(s)$ is the assumed limit of $\sqrt{n}V^n(s)$. That scaled instantaneous loss rate appears as the instantaneous drift term $F'(0)V(s)$ in (3). Replacing that term with the right-hand side of (7) leads to the revised stochastic differential equation

$$\begin{array}{l} V(t) = x + \sigma W(t) + \frac{\theta}{\lambda}t - \int_0^t \left[\int_0^{V(s)} h(u)du\right] ds + L(t) \geq 0 \\ \text{subject to: } L \text{ is non-decreasing, has } L(0) = 0 \text{ and } \int_0^\infty V(s)dL(s) = 0. \end{array} \tag{8}$$

We note that the existence of a unique stationary distribution $V(\infty)$ is always guaranteed. This is because $h(u) \geq 0$ is a hazard rate function, it satisfies $\int_0^\infty h(u)du = \infty$, and therefore, trivially there exists a constant $z_0 \geq 0$ such that $\int_0^z h(u)du > \theta$ for all $z > z_0$.

Theorems 5.1 and 6.1 in Reed and Ward (2008) provide rigourous support (i.e., weak convergence as in (4)) for the approximation of $\sqrt{n}V^n$ by V in (8). This requires the following additional assumption.

($\mathbb{A}4$) The abandonment distribution function F^n and its associated hazard rate function h^n satisfy

$$F^n(V^n(s)) = 1 - \exp\left(\frac{-1}{\sqrt{n}}\int_0^{\sqrt{n}V^n(s)} h^n\left(\frac{v}{\sqrt{n}}\right)dv\right) \text{ and } h^n(x) = h(\sqrt{n}x), \text{ for all } x \geq 0.$$

Remark 1. *(Connecting the Diffusion Approximations (3) and (8)) The diffusion defined by (8) is in some sense a refinement of the one defined by (3). To see this (following intuition provided in Reed and Tezcan (2012) for the $GI/M/N + GI$ model), perform a Taylor series expansion of $h(u)$ about 0 as follows*

$$h(u) = h(0) + \sum_{i=1}^{\infty} \frac{u^i}{i!} h^{(i)}(0).$$

Then, since $h(0) = F'(0)$, it follows that the linear portion of the infinitesimal drift in (3) arises from the lowest order term of the above Taylor series expansion. Hence the simpler diffusion approximation (3) can only be expected to perform well when $h(0)$ is large compared to $h^{(i)}(0)$. This is true for an exponential distribution (since the hazard rate is constant). This is not true for the many Gamma distributions whose density equals zero at the origin. In such cases, the diffusion approximation in (8) is better able to capture the effects of customer abandonments by taking into account the entire *behavior of the abandonment distribution.*

$$\begin{tikzcd}
\sqrt{n}V^n(t) \arrow[r, "t \to \infty"] \arrow[d, "n \to \infty"'] & \sqrt{n}V^n(\infty) \arrow[d, "n \to \infty"] \\
V(t) \arrow[r, "t \to \infty"] & V(\infty)
\end{tikzcd}$$

Figure 1: *A graphical representation of the limit interchange.*

3 The Steady-State Convergence Result

We establish a limit interchange result, represented graphically in Figure 1. Specifically, for V defined by (3) or V defined by (8), it is known that

$$\widetilde{V} \equiv \sqrt{n}V^n \Rightarrow V \quad \text{in} \quad D(\mathbb{R}_+, \mathbb{R}) \quad \text{as} \quad n \to \infty.$$

However, the convergence of the associated steady-state distributions and steady-state moments is not known. For the convergence of the steady-state moments, we require the following slightly higher moment assumption on the primitives.

($\mathbb{A}$5) Assume that

$$\mathbb{E}[u_1^q + v_1^q] < \infty \text{ for some } q > 2.$$

Theorem 1. *Let V be defined by (3) if* ($\mathbb{A}$1)–($\mathbb{A}$3) *hold and by (8) if* ($\mathbb{A}$1)–($\mathbb{A}$4) *hold.*

(a) As $n \to \infty, \widetilde{V}^n(\infty) \Rightarrow V(\infty)$.

(b) If also ($\mathbb{A}$5) *holds, then also*

$$\mathbb{E}[(\widetilde{V}^n(\infty))^m] \to \mathbb{E}[(V(\infty))^m], \quad \textit{as } n \to \infty \tag{9}$$

holds for any $m \in (0, q-1)$, where q is as in ($\mathbb{A}$5).

The proof of Theorem 1 is found in Section 5, after appropriate technical machinery is developed in Section 4.

As a consequence of Theorem 1, we are able to establish the limit behavior of the scaled steady-state abandonment probability and of the scaled queue-length. For the hazard rate scaling case ($\mathbb{A}$4), we impose the following conditions on the hazard rate and moments of the model primitives.

($\mathbb{A}5'$) Assume that

$$h(u) \le K(1+u^l) \text{ for some } l>0 \text{ and } K>0, \text{ for all } u \ge 0,$$

and that

$$\mathbb{E}[u_1^q + v_1^q] < \infty \text{ for some } q \in (2+l, \infty).$$

Corollary 1. *Let $V(\infty)$ be a random variable possessing the steady-state distribution of V in (3). Then, under* ($\mathbb{A}1$)–($\mathbb{A}3$) *and* ($\mathbb{A}5$), *we have*

$$\sqrt{n}P_a^n \to F'(0)\mathbb{E}\left[V(\infty)\right] \quad \textit{and} \quad \mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] \to \lambda\mathbb{E}[V(\infty)], \ \textit{as } n \to \infty. \tag{10}$$

Let $V(\infty)$ be a random variable possessing the steady-state distribution of V in (8). Then, under ($\mathbb{A}1$)–($\mathbb{A}4$) *and* ($\mathbb{A}5'$), *we have*

$$\sqrt{n}P_a^n \to \mathbb{E}\left[\int_0^{V(\infty)} h(u)du\right] \quad \textit{and} \quad \mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] \to \lambda\mathbb{E}[V(\infty)], \ \textit{as } n \to \infty. \tag{11}$$

Proof of Corollary 1. We consider the case that ($\mathbb{A}1$)–($\mathbb{A}5$) hold, and V is defined by (8). The case when ($\mathbb{A}4$) is not assumed is a special case of the former by setting the abandonment distributions $F^n \equiv F$ and $h(u) \equiv F'(0)$ in what follows.

To prove the first assertion in (11), notice that

$$g_n(x) \equiv \sqrt{n}\left(1 - \exp\left(\frac{-x}{\sqrt{n}}\right)\right) \to x,$$

as $n \to \infty$, uniformly on compact sets. This in turn implies that $g_n(x_n) \to x$ whenever $x_n \to x$. From Theorem 1(a) and (generalized) continuous mapping theorem (cf. Theorem 3.4.4 in Whitt (2002), Billingsley (1999)), we have

$$g_n\left(\int_0^{\sqrt{n}V^n(\infty)} h(u)du\right) \Rightarrow \int_0^{V(\infty)} h(u)du.$$

It is straightforward, from the change of variable, to check the left-hand side above after taking an expectation equals to $\sqrt{n}P_a^n \equiv \sqrt{n}\mathbb{E}[F^n(V^n(\infty))]$. From Theorem 1(b) and ($\mathbb{A}5$'), we conclude that $\{\int_0^{\sqrt{n}V^n(\infty)} h(u)du : n \ge 1\}$ is a uniformly integrable family and hence the first assertion follows. The second assertion in (11) is immediate from an asymptotic relationship between the scaled queue-length and offered waiting time processes, $Q^n/\sqrt{n} - \lambda\widetilde{V}^n \Rightarrow 0$ as $n \to \infty$ (see Theorem 6.1 in Reed and Ward (2008)), combined with the weak convergence results in Theorem 1(a) and the converging together theorem. □

4 Uniform Moment Estimates

The main step in proving Theorem 1 is the *tightness* result of the family of steady-state distributions (indexed by $n \geq 1$) of the offered waiting time processes. The key to the desired tightness is to obtain *uniform* (in n) bounds for the moments of the steady-state distributions. Proposition 1 below provides estimates on moments of the (scaled) state process that are uniform in the scaling parameter n.

In this section, it is helpful to emphasize the initial condition of the scaled process $\widetilde{V}^n$. For the initial condition x_0^n of V^n in (1), $\sqrt{n}x_0^n \to x$ as $n \to \infty$, where x is the initial condition of the diffusion V in (8). This is true if, for example, $x_0^n = x/\sqrt{n}$ for a given $x \geq 0$. Without loss of generality, and for notational convenience, we will assume $x_0^n = x/\sqrt{n}$ so that x is the initial condition of $\widetilde{V}^n$, and we will write $\widetilde{V}_x^n$.

We follow the same convention for all processes defined in this section, and use the subscript x to remind the reader that the process definition assumes the initial state of $\widetilde{V}_x^n$ is x.

Proposition 1. *There exist $N_0 \in I\!N$ and $t_0 \in (0, \infty)$ such that for all $t \geq t_0$*

$$\lim_{x\to\infty} \sup_{n \geq N_0} \frac{1}{x^q} I\!E\left[(\widetilde{V}_x^n(tx))^q\right] = 0\,, \tag{12}$$

where $q > 2$ is as in ($\mathbb{A}5$), and $q = 2$ when ($\mathbb{A}5$) is not assumed.

We first prove Proposition 1 and Theorem 1 under the assumption that the hazard rate is scaled (that is, ($\mathbb{A}1$)–($\mathbb{A}4$) hold for part (a) and ($\mathbb{A}5$) additionally holds for part (b)). The case that the hazard rate is not scaled is covered by minor adjustments to the proofs, which are shown at the end of proof of Theorem 1.

The proof of Proposition 1 relies on a martingale representation of the offered waiting time process and the continuity properties of a nonlinear generalized regulator mapping. We first provide this setup, and second give the proof.

Martingale Representation of the Offered Waiting Time Process

Recall that we consider a sequence of systems indexed by $n \geq 1$ in which the arrival rates become large and service times small, and also the system primitives $(t_j^n, v_j^n, d_j^n : j \geq 1)_{n\geq1}$ with t_j^n, v_j^n, d_j^n being the j-th arrival epoch, service time, and abandonment time in the n-th system. Our formulation closely follows that of Reed and Ward (2008).

Define the σ-fields $(\widehat{\mathcal{F}}_i^n)_{i\geq1}$ where

$$\widehat{\mathcal{F}}_i^n \equiv \sigma((t_1^n, v_1^n, d_1^n), \ldots, (t_i^n, v_i^n, d_i^n), t_{i+1}^n) \subseteq \mathcal{F},$$

and let $\widehat{\mathcal{F}}_0^n \equiv \sigma(t_1^n)$. Notice that $V_x^n(t_i^n-)$ is $\widehat{\mathcal{F}}_{i-1}^n$-measurable and the abandonment time d_i^n of the i-th customer is independent of $\widehat{\mathcal{F}}_{i-1}^n$. Hence,

$$I\!P[V_x^n(t_i^n-) \geq d_i^n | \widehat{\mathcal{F}}_{i-1}^n] = F^n(V_x^n(t_i^n-)), \quad i = 1, 2, \ldots, \tag{13}$$

holds almost surely, where F^n is the distribution function of d_i^n. We then have a martingale with respect to the filtration $(\widehat{\mathcal{F}}_i^n)_{i\geq 1}$ given by

$$M_{d,x}^n(i) \equiv \sum_{j=1}^{i} \left(\mathbf{1}_{[V_x^n(t_j^n-)\geq d_j^n]} - I\!E(\mathbf{1}_{[V_x^n(t_j^n-)\geq d_j^n]} | \widehat{\mathcal{F}}_{j-1}^n) \right). \tag{14}$$

Using (13), we also see that for all $i \in I\!N$

$$M_{d,x}^n(i) = \sum_{j=1}^{i} \left[\mathbf{1}_{[V_x^n(t_j^n-)\geq d_j^n]} - F^n(V_x^n(t_j^n-)) \right]. \tag{15}$$

Next, consider the following centered quantities:

$$S^n(i) \equiv \frac{1}{n} \sum_{j=1}^{i} (v_j - 1), \quad S_{d,x}^n(i) \equiv \frac{1}{n} \sum_{j=1}^{i} (v_j - 1) \mathbf{1}_{\{V_x^n(t_j^n-)\geq d_j^n\}}. \tag{16}$$

Algebra from (1)–(2) and (13)–(16) shows that

$$V_x^n(t) = X_x^n(t) + \epsilon_x^n(t) - \int_0^t \left(\int_0^{V_x^n(s-)} h^n(u) du \right) ds + I_x^n(t), \tag{17}$$

where

$$\begin{aligned}
X_x^n(t) &\equiv \frac{x}{\sqrt{n}} + \frac{1}{\mu^n} \left(A^n(t) - \lambda^n t \right) + \left(\frac{n}{\mu^n} \right) \left(S^n\left(A^n(t) \right) - S_{d,x}^n(A^n(t)) - \frac{1}{n} M_{d,x}^n(A^n(t)) \right) + \left(\frac{\lambda^n}{\mu^n} - 1 \right) t, \\
\epsilon_x^n(t) &\equiv \int_0^t \left(\int_0^{V_x^n(s-)} h^n(u) du \right) ds - \frac{1}{\mu^n} \int_0^t F^n(V_x^n(s-)) dA^n(s), \\
I_x^n(t) &\equiv \int_0^t \mathbf{1}_{[V_x^n(s)=0]} ds\,.
\end{aligned}$$

Regulator Mapping Representation

We define fluid-scaled and diffusion-scaled quantities to carry out our analysis. Let

$$\bar{A}^n(t) \equiv \frac{A^n(t)}{n}, \quad \widetilde{A}^n(t) \equiv \sqrt{n} \left(\frac{1}{n} A^n(t) - \frac{1}{n} \lambda^n t \right), \quad \text{and also}$$

$$\widetilde{S}^n(t) \equiv \sqrt{n} S^n([nt]), \quad \widetilde{S}^n_{d,x}(t) \equiv \sqrt{n} S^n_d([nt]), \quad \widetilde{M}^n_{d,x}(t) \equiv \frac{1}{\sqrt{n}} M^n_{d,x}([nt]),$$
$$\widetilde{V}^n_x(t) \equiv \sqrt{n} V^n_x(t), \quad \widetilde{X}^n_x(t) \equiv \sqrt{n} X^n_x(t), \quad \widetilde{\epsilon}^n_x(t) \equiv \sqrt{n} \epsilon^n_x(t), \quad \widetilde{I}^n_x(t) \equiv \sqrt{n} I^n_x(t).$$

The diffusion-scaled offered waiting time process can be written in terms of the fluid- and diffusion-scaled quantities as:

$$\widetilde{V}^n_x(t) = \widetilde{X}^n_x(t) + \widetilde{\epsilon}^n_x(t) - \int_0^t \left(\int_0^{\widetilde{V}^n_x(s-)} h(v) dv \right) ds + \widetilde{I}^n_x(t),$$

where

$$\begin{aligned} \widetilde{X}^n_x(t) &= x + \left(\frac{n}{\mu^n}\right) \left(\widetilde{S}^n(\bar{A}^n(t)) - \widetilde{S}^n_{d,x}(\bar{A}^n(t)) - \widetilde{M}^n_{d,x}(\bar{A}^n(t)) + \widetilde{A}^n(t) \right) + \left(\frac{n}{\mu^n}\right) \sqrt{n} \left(\frac{\lambda^n}{n} - \frac{\mu^n}{n} \right) t, \\ \widetilde{\epsilon}^n_x(t) &= \int_0^t \left(\int_0^{\widetilde{V}^n_x(s^-)} h(v) dv \right) ds - \frac{1}{\mu^n} \int_0^t \frac{1}{\sqrt{n}} F^n \left(\frac{1}{\sqrt{n}} \widetilde{V}^n_x(s^-) \right) d\bar{A}^n(s)\,. \end{aligned}$$

It is straightforward to check using the definition of the idle time process that the process $(\widetilde{V}^n_x, \widetilde{I}^n_x)$ satisfies the conditions for the one-sided nonlinear generalized regulator mapping (cf. Definition 4.1 of Reed and Ward (2008)) $(\phi^h, \psi^h) : D([0,\infty), I\!R) \to D([0,\infty), [0,\infty) \times [0,\infty))$ is defined by $(\phi^h, \psi^h)(y) \equiv (z, l)$ where

(a) $z(t) = y(t) - \int_0^t \left(\int_0^{z(s)} h(u) du \right) ds + l(t) \in [0, \infty)$ for all $t \geq 0$;

(b) l is nondecreasing, $l(0) = 0$, and $\int_0^\infty z(t) dl(t) = 0$.

Hence

$$(\widetilde{V}^n_x, \widetilde{I}^n_x) = (\phi^h, \psi^h)(\widetilde{X}^n_x + \widetilde{\epsilon}^n_x)\,. \tag{18}$$

The proof of Proposition 1 is inspired by Budhiraja and Lee (2009). We expand their approach to incorporate customer abandonments. The key differences are (a) the use of the nonlinear generalized regulator mapping to connect results on the underlying arrival and service renewal processes to results on the offered waiting time process, (b) the use of basic properties of the underlying hazard rate function, and (c) a further consideration of the scaled error process $\widetilde{\epsilon}^n_x$.

Proof of Proposition 1. Let

$$\begin{aligned} \widetilde{N}^n_x(t) &\equiv \left(\frac{n}{\mu^n}\right) \left(\widetilde{A}^n(t) + \widetilde{S}^n(\bar{A}^n(t)) - \widetilde{S}^n_{d,x}(\bar{A}^n(t)) - \widetilde{M}^n_{d,x}(\bar{A}^n(t)) \right) \\ b^n &\equiv \frac{n}{\mu^n} n^{-1/2} (\lambda^n - \mu^n), \end{aligned}$$

so that $\tilde{X}^n_x$ can be succinctly written as

$$\widetilde{X}^n_x(t) = x + b^n t + \widetilde{N}^n_x(t). \tag{19}$$

Recall that b^n converges to θ/λ as $n \to \infty$ from ($\mathbb{A}2$). For the martingale $\widetilde{N}^n_x$ in (19), we note from Section 5.1 of Reed and Ward (2008) that both $\widetilde{S}^n_{d,x}(\cdot)$ and $\widetilde{M}^n_{d,x}(\cdot)$ weakly converge to zero process as $n \to \infty$, and also for all $T \geq 0$, $\lim_{n\to\infty} \sup_{t\in[0,T]} |\bar{A}^n(t) - \lambda t| = 0$ almost surely. These facts, together with the functional central limit theorem $(\widetilde{A}^n, \widetilde{S}^n \circ \bar{A}_n) \Rightarrow (\sqrt{\lambda^3 \operatorname{var}(u_1)} B_1, \sqrt{\lambda \operatorname{var}(v_1)} B_2)$ imply that $\widetilde{X}^n$ weakly converges to a Brownian motion with infinitesimal mean θ/λ and variance $\sigma^2 \equiv \lambda \operatorname{var}(u_1) + \lambda^{-1} \operatorname{var}(v_1)$.

For $n \geq 1$, fix $x \in [0, \infty)$ and write $\widetilde{V}^n_x$ (recall (18) and (19)) as

$$\widetilde{V}^n_x(t) = \phi^h\left(x + b^n i + \widetilde{N}^n_x(\cdot) + \widetilde{\epsilon}^n_x(\cdot)\right)(t), \quad t \geq 0, \tag{20}$$

where $i : [0, \infty) \to [0, \infty)$ is an identity map, i.e., $i(t) \equiv t$ for all $t \in [0, \infty)$. Define a deterministic dynamical system analogous to (20) as

$$Z^n_x(t) \equiv \phi^h\left(x + b^n i\right)(t), \quad t \geq 0. \tag{21}$$

Then, using the Lipschitz property of $\phi^h(\cdot)$ (cf. Proposition 4.1 of Reed and Ward (2008)), we have for some $\kappa \in (0, \infty)$ that

$$|\widetilde{V}^n_x(t) - Z^n_x(t)| \leq \kappa \sup_{0\leq s\leq t} |\widetilde{N}^n_x(s) + \widetilde{\epsilon}^n_x(s)|, \quad \text{for all } t \geq 0. \tag{22}$$

Next, denote by

$$b^*_n(z) \equiv b^n - \int_0^z h(u)du \quad \text{for} \quad z \geq 0.$$

We claim that there exists a constant $D > 0$ such that $Z^n_x(t) = 0$ for all $t \geq Dx$. Notice that $\frac{d}{dz}(b^*_n(z)) = -h(z) \leq 0$, then from ($\mathbb{A}4$) we conclude that there exist $\bar{z}, \delta \in (0, \infty)$ and $N_1 \in I\!N$ satisfying

$$\inf_{n\geq N_1} \inf_{z\in[\bar{z},\infty)} |b^*_n(z)| \geq \delta. \tag{23}$$

The above condition (23) resembles the one-dimensional version of the so-called 'cone condition' in Atar et al. (2001) that was used to characterize stability of a family of diffusion models with state dependent coefficients, constrained to take values in some convex polyhedral cone in $I\!R^d$. The condition (23) says that the state-dependent drift $(b^*_n(z), n \geq N_1, z \geq \bar{z})$ stays away from zero (to the negative direction), *uniformly* both in the parameter n and the state variable z. Thus, for all $n \geq N_1$ and $z \geq \bar{z}$,

$$b^*_n(z) \in \mathcal{C}(\delta) \equiv \{v \in [0, \infty) : |v| \geq \delta\}. \tag{24}$$

For $q_0 \in [0,\infty)$, denote by $\mathcal{T}(q_0)$ the collection of all trajectories of the form

$$\phi(t) = q_0 + \int_0^t \xi(s)ds + l(t) \in [0,\infty), \quad \text{for all } t \geq 0,$$

where $l(\cdot)$ is nondecreasing, $l(0)=0$, $\int_0^\infty \phi(t)dl(t)=0$, and $\xi : [0,\infty) \to I\!\!R$ is a measurable map satisfying for all $t \in [0,\infty)$, $\int_0^t |\xi(s)|ds < \infty$, and $\xi(t) \in \mathcal{C}(\delta)$. Define the hitting time to the origin function

$$T(q_0) \equiv \sup_{\phi \in \mathcal{T}(q_0)} \inf\{t \geq 0 : \phi(t) = 0\}.$$

Then, Lemma 3.1 of Atar et al. (2001) (in particular, the estimate (3.10) therein) shows that

$$T(q_0) \leq \frac{4\kappa^2}{\delta} q_0 \text{ and for all } \phi \in \mathcal{T}(q_0), \quad \phi(t) = 0 \quad \text{for all } t \geq T(q_0),$$

where $\kappa \in (0,\infty)$ is as in (22). Combining this observation with (24), we now see that $Z_x^n(t) = 0$ for all $n \geq N_1$, $t \geq Dx$, where $D = 4\kappa^2/\delta$. Using this estimate in (22), we now have that

$$|\widetilde{V}_x^n(tx)| \leq \kappa \sup_{0 \leq s \leq tx} |\widetilde{N}_x^n(s) + \widetilde{\epsilon}_x^n(s)| \tag{25}$$

for all $n \geq N_1$, $t \geq D$ and for all initial conditions $x \in [0,\infty)$.

Next, we obtain an estimate on the q-th moment ($q \geq 2$) of the right-hand side of (25). For the martingale $\widetilde{N}_x^n$ in (25), it follows from the discussion below (19) that there exists $N \in I\!\!N$ such that for all $n \geq N$,

$$I\!\!E\left[\sup_{0 \leq s \leq t} |\widetilde{N}_x^n(s)|^2\right] \leq c_0(1+t) \quad \text{for some } c_0 \in (0,\infty). \tag{26}$$

(Note that such an estimate is available under the second moment condition $I\!\!E(u_1+v_1)^2 < \infty$.) Moreover, under the q-th moment ($q > 2$) assumption on the primitives as in ($\mathbb{A}$5), we have from standard estimates for renewal processes (see, e.g., Theorem 4 in Krichagina and Taksar (1992)) and observations leading to (26) that there exists $\bar{N} \in I\!\!N$ such that for all $n \geq \bar{N}$,

$$I\!\!E\left[\sup_{0 \leq s \leq t} |\widetilde{N}_x^n(s)|^q\right] \leq c_1(1+t^{1/2})^q \quad \text{for some } c_1 \in (0,\infty). \tag{27}$$

Also, since $\widetilde{\epsilon}_x^n(\cdot) \Rightarrow 0$ uniformly on compact sets as $n \to \infty$ (see Section 5.1.2 of Reed and Ward (2008), and note that our assumed non-zero initial condition is immaterial), there exists an $N_2 \in I\!\!N$ such that for all $n \geq N_2$,

$$I\!\!E\left[\sup_{0 \leq s \leq t} |\widetilde{\epsilon}_x^n(s)|^q\right] \leq c_2 \quad \text{for some } c_2 \in (0,\infty). \tag{28}$$

Applying the estimates in (26), (27) and (28), we now have that for all $q \geq 2$, $t \geq D$ and $x \in [0, \infty)$, and $n \geq N_0 \equiv \max\{\bar{N}, N_1, N_2\}$,

$$\mathbb{E}|\widetilde{V}_x^n(tx)|^q \leq c_3(1 + (tx)^{1/2})^q \quad \text{for some } c_3 \in (0, \infty). \tag{29}$$

Then the desired result (12) follows on choosing $t_0 = D$. □

5 The Steady-State Convergence Proof

We begin by providing a general statement concerning Markov processes which we require to apply in the steady-state convergence proof. For $\bar{\delta} \in (0, \infty)$, define the return time to a compact set $C \subset [0, \infty)$ by $\tau_C^n(\bar{\delta}) \equiv \inf\{t \geq \bar{\delta} : \widetilde{V}_x^n(t) \in C\}$.

Proposition 2. *(Theorem 3.5 of Budhiraja and Lee (2009), cf. Proposition 5.4 in Dai and Meyn (1995)) Let $f : [0, \infty) \to [0, \infty)$ be a measurable map. Define for $\bar{\delta} \in (0, \infty)$, and a compact set $C \subset [0, \infty)$*

$$G_n(x) \equiv \mathbb{E}\left[\int_0^{\tau_C^n(\bar{\delta})} f(\widetilde{V}_x^n(t))dt\right], \quad x \in [0, \infty).$$

If $\sup_n G_n$ is everywhere finite and uniformly bounded on C, then there exists a constant $\eta \in (0, \infty)$ such that for all $n \in \mathbb{N}$, $t \in [\bar{\delta}, \infty)$, $x \in [0, \infty)$,

$$\frac{1}{t}\mathbb{E}[G_n(\widetilde{V}_x^n(t))] + \frac{1}{t}\int_0^t \mathbb{E}[f(\widetilde{V}_x^n(s))]ds \leq \frac{1}{t}G_n(x) + \eta. \tag{30}$$

Proof of Theorem 1. Part (a): Let π^n and π be the unique steady-state distributions of V^n and V in (8) respectively. Since the process-level convergence $\widetilde{V}^n \Rightarrow V$ as $n \to \infty$ follows from Theorem 5.1 in Reed and Ward (2008), it suffices to establish the tightness of the family $\{\pi^n\}$. To prove that tightness, it suffices to show that there exists a positive integer N such that for all $n \geq N$

$$\int_{[0,\infty)} x\pi^n(dx) \leq \tilde{c}, \tag{31}$$

where $\tilde{c} \in (0, \infty)$ is a constant independent of n.

A key observation from Proposition 1 is that there exists a $\gamma_0 \in (0, \infty)$ such that, for t_0 and N_0 as in that same proposition,

$$\sup_{n \geq N_0} \mathbb{E}(\widetilde{V}_x^n(t_0 x))^q \leq \frac{1}{2}x^q, \text{ for } x \in (\gamma_0, \infty) \text{ and } q = 2. \tag{32}$$

Next, we intend to apply Proposition 2 with

$$\bar{\delta} \equiv t_0\gamma_0, \;\; f(x) \equiv 1 + x^{q-1} \text{ for } x \in [0,\infty), \;\text{ and }\; C \equiv \{x \in [0,\infty) : x \leq \gamma_0\}, \tag{33}$$

where $q = 2$. (In the proof of part (b), q will be greater than 2.)

Suppose we can show that there exists a positive integer N and $\overline{c} \in (0,\infty)$ such that

$$\sup_{n\geq N} \mathbb{E}\left[\int_0^{\tau_C^n(\bar{\delta})} (1 + (\widetilde{V}_x^n(t))^{q-1})dt\right] \leq \bar{c}(1 + x^q), \quad x \in [0,\infty), \tag{34}$$

so that the conditions of Proposition 2 are satisfied. Then, for $x \in [0,\infty)$ and $\eta \in (0,\infty)$ the constant in that same proposition,

$$\Phi_n(x) \equiv \frac{1}{t}G_n(x) - \frac{1}{t}\mathbb{E}[G_n(\widetilde{V}_x^n(t))] \geq \frac{1}{t}\int_0^t \mathbb{E}(f(\widetilde{V}_x^n(s)))ds - \eta,$$

for all $n \in \mathbb{N}, t \in [\overline{\delta},\infty)$, and $x \in [0,\infty)$, and so

$$\int_{[0,\infty)} \Phi_n(x)\pi^n(dx) \geq \int_{[0,\infty)} \left(\frac{1}{t}\int_0^t \mathbb{E}(f(\widetilde{V}_x^n(s)))ds - \eta\right)\pi^n(dx). \tag{35}$$

Furthermore, it follows from the definitions of a steady-state distribution π^n and a function $\Phi_n(x)$ that $0 = \int_{[0,\infty)} \Phi_n(x)\pi^n(dx)$ if Φ_n is bounded and measurable, and, otherwise, Fatou's lemma shows that

$$0 \geq \int_{[0,\infty)} \Phi_n(x)\pi^n(dx). \tag{36}$$

Fubini's theorem and the definition of a steady-state distribution show

$$\int_{[0,\infty)} \frac{1}{t}\int_0^t \mathbb{E}(f(\widetilde{V}_x^n(s)))ds\pi^n(dx) = \frac{1}{t}\int_0^t \int_{[0,\infty)} \mathbb{E}(f(\widetilde{V}_x^n(s)))\pi^n(dx)ds = \int_{[0,\infty)} f(x)\pi^n(dx). \tag{37}$$

Finally, it follows from (35)–(37) that

$$0 \geq \int_{[0,\infty)} f(x)\pi^n(dx) - \eta, \tag{38}$$

which establishes (31).

It remains to show (34). Given the estimate in (32), the proof follows along the same lines as those of Theorem 3.4 in Budhiraja and Lee (2009), because $\bar{\delta}$, $f(x)$ and C in (33) are chosen in the same way as in the cited theorem. Following steps analogous to those leading to (39)–(41) in Budhiraja and Lee (2009), it suffices to check that the deterministic trajectory Z_x^n in (21) satisfies the following property: For every $c_1 \in (0,\infty)$, there exist $N \in I\!\!N$ and $c_2 \in (0,\infty)$ such that

$$\sup_{n\geq N} \sup_{0\leq t\leq c_1 x} |Z_x^n(t)| \leq c_2(1+x), \quad x \in [0,\infty). \tag{39}$$

From the Lipschitz continuity of the generalized regulator map ϕ^h, we have

$$\sup_{0\leq t\leq c_1 x} |Z_x^n(t)| \leq \kappa(x + \sup_{0\leq t\leq c_1 x} |b^n t|) \quad \text{for some } \kappa \in (0,\infty).$$

Letting $c_3 \equiv \sup_{n\geq N} |b^n|$ for sufficiently large $N \in I\!\!N$, we have $|b^n t| \leq c_3 t$ since $b^n \to \theta/\lambda \in (-\infty,\infty)$ as $n \to \infty$. Then, the desired estimate (39) follows on setting $c_2 \equiv \kappa + c_1 c_3$. Therefore, the key step (41) in the proof of Theorem 3.4 Budhiraja and Lee (2009) is now satisfied and this completes the proof.

Part (b): Under ($\mathbb{A}$5), Proposition 1 implies (32) with $q > 2$. Therefore, we have a uniform (in n) moment bound on $I\!\!E[\widetilde{V}_x^n(\infty)^{q-1}]$ obtained in (38), with $q > 2$ as in ($\mathbb{A}$5). Combining the weak convergence result in part (a) and a uniform integrability of $\{\widetilde{V}_x^n(\infty)^{q-1}\}_{n\geq 1}$, we conclude the desired moment convergence result.

The proof modifications when the **hazard rate is not scaled are:**

- In the primitives used in Section 4, to obtain the martingale representation of the offered waiting time process, d_j^n and F^n are replaced by d_j and F.
- The definition of ϵ_x^n is replaced by

$$\epsilon_x^n(t) = \int_0^t F'(0) V_x^n(s-)ds - \frac{\lambda^n}{\mu^n}\int_0^t F(V_x^n(s-))d\left(\frac{A^n(s)}{\lambda^n}\right).$$

- The diffusion-scaled offered waiting time process is now represented as:

$$\widetilde{V}_x^n(t) = \widetilde{X}_x^n(t) + \widetilde{\epsilon}_x^n(t) - \int_0^t F'(0) V_x^n(s-)ds + \widetilde{I}_x^n(t).$$

- The function $h(u)$ is replaced by the constant $F'(0)$. Then the nonlinear regulator mapping ϕ^h is going to become a linear regulator mapping, and the definition of b_n^* is going to change so that the second term is linear.

□

Acknowledgements: The research work of Chihoon Lee is supported in part by the Army Research Office (W911NF-14-1-0216).

References

Atar, R., A. Budhiraja, P. Dupuis. 2001. On positive recurrence of constrained diffusion processes. *Ann. Probab.* **29**(2) 979–1000.

Bacelli, F., P. Boyer, G. Hébuterne. 1984. Single-server queues with impatient customers. *Adv. in Appl. Probab.* **16**(4) 887–905.

Billingsley, P. 1999. *Convergence of Probability Measueres, 2nd edition*. John Wiley and Sons, Inc.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. *Advances in queueing*. Probab. Stochastics Ser., CRC, Boca Raton, FL, 463–480.

Budhiraja, A., C. Lee. 2009. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research* **34**(1) 45–56.

Dai, J. G., A. B. Dieker, X. Gao. 2014. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems* **78**(1) 1–29.

Dai, J. G., S. P. Meyn. 1995. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control* **40** 1889–1904.

Gamarnik, D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in open queueing networks. *Ann. Appl. Probab.* **16**(1) 56–90.

Kingman, J. F. C. 1961. The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.* **57** 902–904.

Kingman, J. F. C. 1962. On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* **24** 383–392.

Krichagina, E. V., M. I. Taksar. 1992. Diffusion approximation for $GI/G/1$ controlled queues. *Queueing Systems Theory Appl.* **12**(3-4) 333–367.

Lillo, R. E., M. Martín. 2001. Stability in queues with impatient customers. *Stoch. Models* **17**(3) 375–389.

Mandelbaum, A., P. Momcilovic. 2012. Queue with many servers and impatient customers. *Mathematics of Operations Research* **37**(1) 41–65.

Reed, J. E., T. Tezcan. 2012. Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Operations Research* **60**(4) 981–995.

Reed, J. E., A. R. Ward. 2008. Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. *Mathematics of Operations Research* **33**(3) 606–644.

Ward, A. R., P. W. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43**(1/2) 103–128.

Ward, A. R., P. W. Glynn. 2005. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems* **50**(4) 371–400.

Whitt, W. 2002. *Stochastic-process Limits*. Springer-Verlag, New York.